First steps toward finding relevant pathology-gene pairs using analogy

Marie-Dominique Devignes[1], Yohann Fransot[1], Yves Lepage[2], Jean Lieber[1], Emmanuel Nauer[1], and Malika Smaïl-Tabbone[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
firstname.lastname@loria.fr
[2]IPS, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, 808-0135 Fukuoka-ken, Japan
yves.lepage@waseda.jp

Abstract. This paper presents a first study to infer pathology-gene relation instances using analogy. The meaning of this relation between a pathology $\mathtt{P}$ and a gene $\mathtt{G}$ is "A mutation of $\mathtt{G}$ in a person can cause the appearance of pathology $\mathtt{P}$ for this person." In this work, a pathology is represented by a set of classes from HPO, the Human Phenotype Ontology, whereas a gene is represented similarly, but using GO, the Gene Ontology. Some $(\mathtt{P}, \mathtt{G})$ instances of the pathology-gene relations are known and the idea is to use analogical reasoning to infer new relations. The schema of the inference is as follows: if a target pathology $\mathtt{P}$ is in analogy with three other pathologies $\mathtt{P}_A$, $\mathtt{P}_B$ and $\mathtt{P}_C$ for which associated genes $\mathtt{G}_A$, $\mathtt{G}_B$ and $\mathtt{G}_C$ are known, then it is plausible that the gene $\mathtt{G}$, to be associated with $\mathtt{P}$, is in analogy with $\mathtt{G}_A$, $\mathtt{G}_B$, $\mathtt{G}_C$. This idea has proven to be fruitful in other domains, such as machine translation.
The preliminary question explored in this paper is the following: given four pathologies $\mathtt{P}_A, \mathtt{P}_B, \mathtt{P}_C$ and $\mathtt{P}_D$ in analogy and for which the associated genes $\mathtt{G}_A$, $\mathtt{G}_B$, $\mathtt{G}_C$ and $\mathtt{G}_D$ are known, are thess genes in analogy, or, at least, in approximate analogy?
Results of a large scale analysis ($4,000$ $(\mathtt{P}, \mathtt{G})$ pairs) reveal that the quadruples of genes associated with quadruples of pathologies in analogy do not display statistically different analogical dissimilarity values than randomly selected quadruples of genes. Nevertheless very low analogical dissimilarity values are found in a small subset of gene quadruples that are specifically associated with pathologies in analogy. Analysis of these quadruples may allow us to learn more sophisticated analogical relations on genes in order to improve the recovery of pathology-gene pairs using analogy.

Keywords: analogy, gene-pathology relation, ontology, annotation

1 Introduction

Transposing proportional analogies from a domain to another one is a general principle in problem solving. It has been applied to several problems in natural language processing, like grapheme-phoneme transcription (i.e., pronounciation) [7], morphological analysis [3], syntactic analysis [1] or machine translation [4]. A *proportional analogy* is

a quaternary relation between four objects A, B, C and D denoted by $A : B :: C : D$. It is read: "A is to B as C is to D". As for the special case of translation, the approach is based on a set of cases, a case being an ordered pair (S, S') where S is a sentence in a first language and S' is a translation of S in a second language. The principle consists in, given a sentence D in the first language, (1) finding three cases (A, A'), (B, B'), (C, C') verifying $A : B :: C : D$ in the first language and (2) solving the *analogical equation* $A' : B' :: C' : x$ in the second language. The following example illustrates this idea, for translation from French into English:

$$\begin{aligned}
(A, A') &= (\textit{Tu évites de danser le tango ?}, && \textit{Do you avoid dancing tango?})\\
(B, B') &= (\textit{J'évite de manger du melon.}, && \textit{I avoid eating melon.})\\
(C, C') &= (\textit{Tu aimes danser le tango ?}, && \textit{Do you like dancing tango?})\\
D &= \textit{J'aime manger du melon.} && \text{(target problem)}
\end{aligned}$$

As the relation $A : B :: C : D$ holds, the three cases are returned and it is inferred that a candidate translation D' of D is a solution of the analogical equation $A' : B' :: C' : x$. In this example, the solution $x = \textit{I like eating melon.}$ is a correct translation D' of D. In [5], this approach has been reformulated in case-based reasoning terms and some extensions are proposed.

Now, the question is whether this approach can apply to different data than language data. This paper examines this issue in the context of pathology-gene pairs $(\mathtt{P}, \mathtt{G})$: roughly said, such a case means that the gene $\mathtt{G}$ plays an important role in the pathology $\mathtt{P}$. Thus the idea is to use proportional analogies to mine a pathology-gene base $\mathcal{B}$, in order to find hypotheses of new pairs $(\mathtt{P}, \mathtt{G})$, according to the following inference rule:

$$\frac{\begin{array}{c}\mathtt{P}_A : \mathtt{P}_B :: \mathtt{P}_C : \mathtt{P}_D \quad \text{in the pathology space}\\ (\mathtt{P}_A, \mathtt{G}_A), (\mathtt{P}_B, \mathtt{G}_B), (\mathtt{P}_C, \mathtt{G}_C) \in \mathcal{B}\\ x \text{ is a solution of } \mathtt{G}_A : \mathtt{G}_B :: \mathtt{G}_C : x \quad \text{in the gene space}\end{array}}{\text{It is plausible that } (\mathtt{P}_D, x) \text{ is a relevant pathology-gene pair}} \tag{1}$$

To address this issue, some notions related to proportional analogy are necessary: they are introduced in Section 2. It is also necessary to have some biological notions about pathologies and genes, in particular about their representations: the goal of Section 3 is to introduce these notions. Based on all these notions, the first approach to examine this inference in this application domain is detailed in Section 4 and the implementation principles are presented in Section 5. Section 6 presents the first results and interpretations. Finally, Section 7 concludes and proposes several research directions..

2 Proportional Analogy : Definitions

Basic Definitions. Let $\mathcal{U}$ be a set. A *proportional analogy* on $\mathcal{U}$ is a quaternary relation on members of $\mathcal{U}$ that is usually denoted as $A : B :: C : D$. It is read: "A is to B as C is to D". In this expression, $A : B$ and $C : D$ are called *ratios*, and the binary relation :: is called *conformity*. A proportional analogy (PA) generally satisfies the following postulates: for any $(A, B, C, D) \in \mathcal{U}^4$,

– $A : B :: A : B$ is always true (reflexivity of conformity);
– if $A : B :: C : D$ then $C : D :: A : B$ (symmetry of conformity);
– if $A : B :: C : D$ then $A : C :: B : D$ (exchange of the means).

Other properties, like the one called *exchange of the extremes*, can be deduced from the previous postulates: if $A : B :: C : D$ then $D : B :: C : A$. Indeed, for one given analogy, there exist seven other equivalent forms.

The reflexivity of conformity gives birth to many analogies $A : B :: A : B$, equivalent to $A : A :: B : B$, by exchange of the means. Such analogies are poorly informative in that they give no information about A and B or about their relation. In the sequel, we call them *flat analogies*.

Boolean Representations. Given $\mathcal{U} = \mathbb{B} = \{0, 1\}$ the set of Booleans (where 0 and 1 represent `false` and `true`, respectively), the proportional analogy defined by $A : B :: C : D$ if $B - A = D - C$ (where these differences belong to $\{-1, 0, 1\}$) satisfies the PA postulates. There are six patterns $ABCD$ satisfying this relation: 0000, 1111, 0011, 1100, 0101 and 1010.

In $\mathcal{U} = \mathbb{B}^n$, the relation defined by a component by component analogy — i.e., $A : B :: C : D$ if $a_i : b_i :: c_i : d_i$ for every $i \in \{1, 2, \ldots, n\}$ — also satisfies this postulates (where, e.g., $A = (a_1, a_2, \ldots, a_n)$). For example,

$$(0, 1, 1, 0, 0) : (1, 1, 0, 1, 0) :: (0, 0, 1, 0, 0) : (1, 0, 0, 1, 0) \tag{2}$$

An n-tuple of Booleans $T = (T_1, T_2, \ldots, T_n)$ can be encoded by the set $\widehat{T}$ of its indices with the value 1. For example, if $T = (0, 1, 1, 0, 1, 0)$, then $\widehat{T} = \{2, 3, 5\}$. Now, given T, $U \in \mathbb{B}^n$, let $\texttt{key}(T, U) = (\widehat{T} \setminus \widehat{U}, \widehat{U} \setminus \widehat{T})$. In fact $\texttt{key}(T, U)$ encodes the changes from T to U. It can be shown that the proportional analogy on $\mathcal{U} = \mathbb{B}^n$ defined above can be characterized by:

$$A : B :: C : D \quad \text{iff} \quad \texttt{key}(A, B) = \texttt{key}(C, D) \tag{3}$$

This can be verified on the example (2) as:

$$\texttt{key}(A, B) = \texttt{key}(C, D) = (\{3\}, \{1, 4\})$$

When, in the data, the n-tuples of Booleans are sparse (i.e., they contain a majority of 0), the interest in this characterization is algorithmic: the size necessary to encode $\texttt{key}(A, B)$ is much smaller than n.

Analogical Dissimilarity. If four objects $(A, B, C, D) \in \mathcal{U}^4$ are *not* in proportional analogy, a question that can be raised is "How far are these objects from forming an analogy?" An analogical dissimilarity (AD) [6] is a function $\texttt{AD} : (A, B, C, D) \mapsto [0, +\infty[$ satisfying the following postulates: for any $(A, B, C, D, E, F) \in \mathcal{U}^6$,

– $\texttt{AD}(A, B, C, D) = 0$ iff $A : B :: C : D$ (consistency with analogy);
– $\texttt{AD}(A, B, C, D) = \texttt{AD}(C, D, B, A)$ (symmetry, reflecting symmetry of conformity);
– $\texttt{AD}(A, B, C, D) = \texttt{AD}(A, C, B, D)$ (central permutation, i.e., exchange of the means);

- $\texttt{AD}(A, B, E, F) \leq \texttt{AD}(A, B, C, D) + \texttt{AD}(C, D, E, F)$ (triangle inequality);
- If $A \neq B$ then $\texttt{AD}(A, B, C, D) \neq \texttt{AD}(B, A, C, D)$.

In $\mathcal{U} = \mathbb{B}$, AD defined by $\texttt{AD}(A, B, C, D) = |(B - A) - (D - C)| \in \{0, 1, 2\}$ for $A, B, C, D \in \mathbb{B}$ satisfies the AD postulates.

In $\mathcal{U} = \mathbb{B}^n$, AD defined by $\texttt{AD}(A, B, C, D) = \sum_{i=1}^{n} \texttt{AD}(A_i, B_i, C_i, D_i)$ also satisfies the AD postulates. For example,

$$\texttt{AD}((0,0,0,0,1),(1,0,1,1,0),(0,1,0,1,0),(1,0,0,0,1)) = 0 + 1 + 1 + 2 + 2 = 6$$

3 Pathology-Gene Relations

A gene is a sequence of nucleotides along a segment of DNA that encodes instructions for RNA synthesis, which, when translated into protein, leads to the expression of phenotypes. The genes are thus the basic physical units of heredity. Phenotypes such as pathologies (or diseases) are associated to genes in some public databases. The most famous one is the OMIM database which focuses on human hereditary diseases (`http://omim.org/`). OMIM provides pathology-gene relations that were carefuly curated and documented w.r.t. the literature. Today, a large number of diseases lack responsible gene(s) hence the numerous gene prioritization methods [2].

On one hand, genes are annotated with their known functions (in fact the functions are accomplished by the proteins produced by the genes). Such functions are taken from GO (Gene Ontology), an ontology encompassing thousands of terms (or classes) mainly linked with subclass-of relations (`http://www.geneontology.org/`). The GO is stuctured as an r-DAG (rooted directed acyclic graph). Some of the gene-GO term relations are based on experimental evidence or published papers whereas others are simply inferred from known relationships combined with gene sequence similarity for instance. The gene-GO term relations are indeed qualified with evidence codes. Manually-assigned evidence codes fall into four general categories: experimental (such as EXP: inferred from experiment), computational analysis (e.g., ISS: inferred from sequence or structural similarity), author statements (for instance TAS: traceable author statement), and curatorial statements (for example IC: inferred by curator). Only one evidence code (IEA: inferred from electronic annotation) is not assigned by a curator. Such gene annotations in many species are available in a public database, AMIGO (`http://amigo.geneontology.org/amigo`).

On the other hand, diseases are annotated with their known associated phenotypes (a.k.a. symptoms) taken from HPO (Human Phenotype Ontology). Similarly to genes and GO-terms, HPO is structured as a r-DAG and disease-HPO term (or class) relations are qualified with evidence codes (e.g. PCS for published clinical study, ICE for individual clinical experience, ITM for inferred from text mining, TAS and IEA having the same meaning as for gene-GO relations) depending on the origin of the relationship. Disease annotations are also stored in the OMIM database.

Table 1 shows an example of disease-gene relationship along with their respective annotations.

Pathologie or Gene	Name	Abbr.	HPO or GO term
Pathologie	Cardiomyopathy dilated 1I	CMD1I	Reduced systolic function (TAS) Congestive heart failure (TAS)
Gene	Desmin	DES	Muscle contraction (TAS) Regulation of heart contraction (TAS) Structural constituent of cytoskeleton (TAS) Intermediate filament (IEA)

Table 1: Example of $(\mathtt{P}, \mathtt{G})$ pair with associated HPO and GO annotations. Evidence codes are between parentheses.

In this work, a pathology $\mathtt{P}$ is represented as a tuple of Booleans in the following way. Let m be the number of classes of HPO and $\{\mathtt{CP}_1, \mathtt{CP}_2, \ldots, \mathtt{CP}_m\}$ be the set of these classes. $\mathtt{P}$ is described by the m-tuple $(p_1, p_2, \ldots, p_m) \in \mathbb{B}^m$ such that $p_i = 1$ iff $\mathtt{P}$ is described by $\mathtt{CP}_i$ (for each $i \in \{1, 2, \ldots, m\}$). It must be noted that if a class $\mathtt{CP}_i$ is a subclass of $\mathtt{CP}_j$ (either directly or by transitivity of the subclass-of relation), then a pathology $\mathtt{P}$ described by $\mathtt{CP}_i$ is also described by $\mathtt{CP}_j$: if $p_i = 1$ then $p_j = 1$ (this is mere application of the deductive closure based on the subclass-of relation). The tuple $(p_1, p_2, \ldots, p_n)$ is sparse: only a small proportion of the classes of HPO are used to describe each single pathology.

The representation of a gene $\mathtt{G}$ is done in a similar way by a tuple $(g_1, g_2, \ldots, g_n) \in \mathbb{B}^n$, where n in the number of classes in GO.

4 Proposed Approach

The goal of the proposed approach is to examine whether the inference rule (1) that associates to a pathology $\mathtt{P}_D$ a gene $\mathtt{G}_D$ gives a relevant pathology-gene pair $(\mathtt{P}_D, \mathtt{G}_D)$, at least with a reasonable frequency. If, e.g., the answer was that it holds 10% of the time, it would still be interesting in a knowledge discovery perspective: providing an expert with original hypotheses with such a proportion of correctness still remains interesting. Having this is mind, two experiments were conducted.

First experiment. The set of quadruples $(\mathtt{P}_A, \mathtt{P}_B, \mathtt{P}_C, \mathtt{P}_D)$ of pathologies (in the chosen representation formalism) such that $\mathtt{P}_A : \mathtt{P}_B :: \mathtt{P}_C : \mathtt{P}_D$ is computed. Only non flat analogies were kept. A gene is associated to each pathology: $\mathtt{G}_A$, $\mathtt{G}_B$, $\mathtt{G}_C$ and $\mathtt{G}_D$. The experiment is designed to meet two objectives:

- find out the proportion of quadruples of genes (related to analogies on pathologies) which are in analogy;
- the situation where $\mathtt{P}_A : \mathtt{P}_B :: \mathtt{P}_C : \mathtt{P}_D$ holds but $\mathtt{G}_A : \mathtt{G}_B :: \mathtt{G}_C : \mathtt{G}_D$ does not hold, may still be interesting if the analogical dissimilarity between these genes is low (i.e., it is close to an exact analogy). Thus, the second objective is to compare the distribution of $\mathtt{AD}(\mathtt{G}_A, \mathtt{G}_B, \mathtt{G}_C, \mathtt{G}_D)$ provided that the corresponding pathologies are

in analogy with the distribution of $\mathtt{AD}(\mathtt{G}_A, \mathtt{G}_B, \mathtt{G}_C, \mathtt{G}_D)$ in general. In particular if the average of the first distribution is significatively lower than the average of the second distribution, this would mean that analogies between genes are somehow connected with analogies between pathologies.

Second Experiment. The second experiment examines the inference the other way round: from genes to pathologies, i.e., the following inference

$$\frac{\begin{array}{c}\mathtt{G}_A : \mathtt{G}_B :: \mathtt{G}_C : \mathtt{G}_D \quad \text{in the gene space} \\ (\mathtt{P}_A, \mathtt{G}_A), (\mathtt{P}_B, \mathtt{G}_B), (\mathtt{P}_C, \mathtt{G}_C) \in \mathcal{B} \\ x \text{ is a solution of } \quad \mathtt{P}_A : \mathtt{P}_B :: \mathtt{P}_C : x \quad \text{in the pathology space}\end{array}}{\text{It is plausible that } (\mathtt{P}_D, x) \text{ is a relevant pathology-gene pair}} \tag{4}$$

The same objectives as in the first experiment are pursued in the reverse direction.

5 Implementation Principles

The main steps of the approach are schematized on Figure 1. All of the data has been loaded into a database and further expanded within the database by (i) deductive closure on HPO and GO annotations and (ii) computation of keys for each pathology (respectively gene) pair.

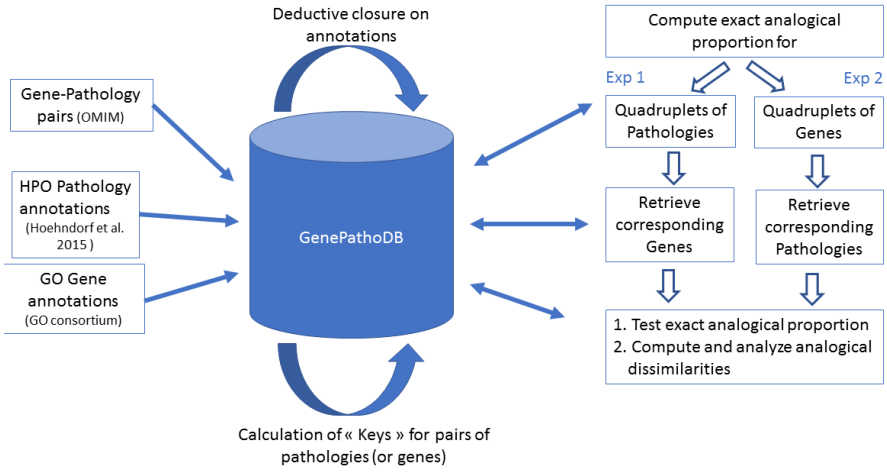

Fig. 1: Outline of the general approach for finding relevant pathology-gene pairs using analogy.

The main algorithmic difficulty was to efficiently find those quadruples $(\mathtt{P}_A, \mathtt{P}_B, \mathtt{P}_C, \mathtt{P}_D)$ of pathologies that are in analogy: the naïve algorithm is in $O(n^4)$,

where n is the number of pathologies for which a gene is known. Since $n \simeq 5000$, this solution is intractable.

The idea is to use the characterization (3) of analogies based on keys. First, the data on pathologies and genes where stored in tables. A table for pathology pairs $(\mathtt{P}_A, \mathtt{P}_B)$ with their keys $\mathtt{key}(\mathtt{P}_A, \mathtt{P}_B)$ was built. A query on this table with a `GROUP BY` clause on these keys is executed. When there are several lines in a group, they correspond to two pairs $(\mathtt{P}_A, \mathtt{P}_B)$ and $(\mathtt{P}_C, \mathtt{P}_D)$ such that $\mathtt{P}_A : \mathtt{P}_B :: \mathtt{P}_C : \mathtt{P}_D$. The flat analogies are subsequently removed.

Other algorithmic difficulties were overcome in a similar way.

6 First Results

The results of the two experiments described in Section 4 are presented below.

For the first experiment, The number of quadruples of pathologies that are non flat analogies is $3,501$ in the chosen representation. Unfortunately, there is no quadruple of genes (related to the retrieved quadruples of pathologies) which are in analogy.

As for the second objective, a computation of $\mathtt{AD}(\mathtt{G}_A, \mathtt{G}_B, \mathtt{G}_C, \mathtt{G}_D)$ for about $3,500$ quadruples of genes chosen at random has given the following result:

$$\text{mean} = 240 \qquad \text{median} = 222 \qquad \text{standard deviation} = 111$$

For the $3,501$ quadruples of genes associated to quadruples of pathologies that are in non flat analogies, the computation of their analogical dissimilarity has given a distribution with the following features:

$$\text{mean} = 247 \qquad \text{median} = 230 \qquad \text{standard deviation} = 112$$

These basic statistics are quite similar, suggesting that in average the distributions are similar between randomly selected quadruples of genes and quadruples of genes associated with quadruples of pathologies in analogy. However close inspection of the distribution histogram reveals that a small subset of these gene quadruples (associated with pathologies in analogy) display very low values of analogical dissimilarity not reached by randomly selected gene quadruples. These quadruples are currently under investigation by experts.

The results in the second experiment, which takes the inference rule the other way round (from gene to pathology), are similar to those in the first experiment.

7 Conclusion and Future Work

This paper examined the following hypothesis: "If four pathologies are in analogy, is it plausible that the corresponding genes are in analogy?" With our databases, under the representation choices of pathologies and genes we adopted and within the simple proportional analogy and analogical dissimilarity framework presented in this paper, the answer is negative for the great majority of pathologies in analogy.

The results we obtained on a large scale analysis ($4,000$ $(\mathtt{P}, \mathtt{G})$ pairs) reveal that the quadruples of genes associated with quadruples of pathologies in analogy do not

display statistically different analogical dissimilarity values compared to randomly selected quadruples of genes. Nevertheless very low analogical dissimilarity values are found in a small subset of gene quadruples that are specifically associated with pathologies in analogy. Analysis of these quadruples may allow us to learn more sophisticated analogical relations on genes in order to improve the recovery of pathology-gene pairs using analogy.

Therefore, an idea for future work would be to find analogical relations on pathologies and on genes so that the inference (1) from pathologies to genes or the inference (4) from genes to pathologies work better. A way to do it would be to keep the analogical relation between pathologies defined in this paper and *learn* the analogical relation between genes. More precisely, let $\mathcal{A}_\lambda$ be an analogical proportion between genes parameterized by λ (that can be, e.g., a tuple of integers). The training set would be quadruples of genes $(\mathtt{G}_A, \mathtt{G}_B, \mathtt{G}_C, \mathtt{G}_D)$ that correspond to pathologies that are in analogy. The objective of the learning method would be to find λ such that an important proportion of the quadruples in the training set are in analogy according to $\mathcal{A}_\lambda$. Once this learning process is achieved, it is hoped that the analogical inference described by (1) with the classical analogy relation between pathologies and the analogy relation $\mathcal{A}_\lambda$ between genes provides an efficient way to mine pathology-gene pairs.

References

1. Ando, S.I., Lepage, Y.: Linguistic structure analysis by analogy: Its efficiency. In: Proceedings of NLPRS-97, Phuket (December 1997) 401–406
2. Driel, M.A.V., Brunner, H.G.: Bioinformatics methods for identifying candidate disease genes. Hum. Genomics (2006) 429–432
3. Lavallée, J.F., Langlais, P.: Unsupervised morphological analysis by formal analogy. In: Lecture Notes in Computer Science. (2010) 8 pages
4. Lepage, Y., Denoual, É.: Purest ever example-based machine translation: detailed presentation and assessment. Machine Translation **19** (2005) 251–282
5. Lepage, Y., Lieber, J.: Case-based translation: First steps from a knowledge-light approach based on analogy to a knowledge-intensive one. In: Proceedings of the Computational Analogy Workshop at the 25th International Conference on Case-Based Reasoning (ICCBR-18), Stockholm, Sweden (August 2018)
6. Miclet, L., Bayoudh, S., Delhay, A.: Analogical dissimilarity: Definition, algorithms and two experiments in machine learning. Journal of Artificial Intelligence Research **32** (2008) 793–824
7. Pirelli, V., Federici, S.: "Derivational" paradigms in morphonology. In: Proceedings of COLING-94. Volume I., Kyōto (August 1994) 234–240